5P

УДК 004.89:811.1/.8

Н.К. Рубашко
Белорусский государственный университет, г. Минск, Беларусь
roubashko@bsu.by

Разработка базовых лингвистических ресурсов естественного языка для информационных систем

В статье излагаются основные принципы разработки базовых лингвистических ресурсов естественного языка. Приводится перечень основных составляющих, дается описание технологии создания этих ресурсов. В качестве примера использования предлагаемой технологии рассматривается разработка базовых лингвистических ресурсов белорусского и русского языков.

Введение

В настоящее время формируется новое поколение информационных технологий, основанных на концепции интегрированной информационной среды, обеспечивающей хранение, обработку и распространение значительного объема информации в промышленных масштабах. Это приводит к тому, что меняется подход к анализу естественного языка (ЕЯ) – на первый план выходят вопросы, связанные с разработкой и эффективным использованием моделей представления знаний о ЕЯ (или иначе, лингвистических знаний) в условиях хорошо организованных лингвистических экспериментов.

При решении проблемы разработки и представления знаний о ЕЯ необходимо, прежде всего, учитывать, что язык, «являющийся средством отражения действительности в человеческом коллективе… представляет собой незамкнутую и поэтому не формализуемую до конца систему» [1]. Трудность или даже невозможность полной ее формализации обусловлена следующими свойствами ЕЯ [2], [3]:
– конфронтацией визуального (языкового, «словарного») значения и отличного от него актуального (текстового) смысла лингвистической единицы;
– парадоксом языка и идиолекта (индивидуального владения языком), проявляющегося в различных интерпретациях границ значения слова у отдельных носителей языка;
– постоянной изменчивостью языка как во времени, так и в географическом и социальном пространстве;
– потенциальной бесконечностью и открытостью лингвистических множеств, обусловленных динамичностью и метафоричностью ЕЯ;
– нечеткостью лингвистических объектов (в первую очередь семантических) и размытостью границ совокупностей этих объектов.

Лингвистические знания должны обеспечивать многоаспектность изучения и самые разнообразные преобразования реального языкового материала, как правило, очень большого объема [4]. Можно сказать, что формализованное представление лингвистических знаний есть не что иное, как лингвистические ресурсы интеллектуальной информационной системы. Эти ресурсы в виде совокупности различного рода корпусов текстов, словарей, грамматик, правил построения семантических кон-

струкций, иными словами, в виде так называемых лингвистических баз знаний (ЛБЗ), являются составной частью развитых лингвистических процессоров (ЛП), обеспечивающих создание совершенно новых технологий работы с текстовыми документами, которые включают их автоматическое чтение, речевой ввод/вывод, ЕЯ-интерфейс пользователя, семантический поиск, машинный перевод и автоматическое реферирование, распознавание, извлечение и управление знаниями и т.п.

Каждый модуль информационной системы, выполняющей определенную обработку ЕЯ, можно разделить на алгоритмическую (функциональную) часть и ее лингвистическое наполнение. При проектировании систем лингвистические ресурсы выделяются в отдельный блок и хранятся в формате, доступном для изменения. Это обеспечивает разделение работы экспертов, инженеров по знаниям и программистов, что очень важно при разработке интеллектуальных информационных систем [5].

Работы по созданию лингвистических ресурсов национальных языков ведутся во всех развитых странах мира, поскольку давно стало очевидным, что создание таких ресурсов – это путь к созданию новых информационных технологий, базирующихся на системах обработки данных на ЕЯ в интеллектуальной среде общения человека и компьютера. Например, в рамках созданного в 1992 г. в США лингвистического консорциума (LDC – Linguistic Data Consortium), обеспечивающего механизм координации крупномасштабных исследований и распределения ресурсов в области информационных технологий, ведутся работы практически со всеми языками мира, в том числе с восточно- и центрально-европейскими языками: болгарским, чешским, эстонским, венгерским, румынским, словенским, русским [6].

Целью данной статьи является описание основных принципов и этапов разработки базовых лингвистических ресурсов ЕЯ, использования предлагаемой технологии при разработке базовых лингвистических ресурсов белорусского языка, а также описание приложений разработанных лингвистических ресурсов.

1. Понятие базовых лингвистических ресурсов

Под базовыми лингвистическими ресурсами любого ЕЯ для информационных систем понимаются:
– исходный корпус текстов данного ЕЯ;
– классификатор свойств ЕЯ на различных уровнях его глубины;
– базовый словарь ЕЯ;
– аннотированный корпус текстов данного ЕЯ (иначе называемый эталонным);
– распознающие лингвистические модели анализа текста на различных уровнях глубины ЕЯ.

В создании лингвистических ресурсов сегодня главную роль играют корпусы текстов – некоторые определенным образом подобранные конечные множества текстов языка. Будем эти тексты называть исходным корпусом текстов (ИКТ) заданного языка L.

В компьютерной лингвистике принято следующее определение: корпус текстов – это вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты, включающие полные фрагменты макроструктуры текстов данной проблемной области. Это определение основывается на следующих признаках [7]:
– корпус текстов должен быть достаточно большого объема;
– он должен быть структурированным или размеченным;
– тексты, составляющие определенный корпус, должны храниться в электронном виде;
– в понятие «электронный корпус» входит, как правило, специальное программное обеспечение для работы с этим корпусом.

Значимость корпуса текстов состоит в следующем:

а) однажды созданный корпус может многократно использоваться;

б) корпус показывает языковые данные в их реальном окружении, что позволяет исследовать лексическую и грамматическую структуру языка, а также непрерывные процессы языковых изменений, происходящие в языке на протяжении определенного отрезка времени;

в) корпус характеризуется представительностью, или сбалансированным составом текстов, что позволяет использовать его для тестирования поисковых машин, машинных морфологий, систем перевода и т.п., а также использовать его в различных лингвистических исследованиях.

Первая задача при создании корпуса состоит в определении его объема, поскольку частотность и релевантность любого лингвистического явления прослеживается тем лучше, чем больше словоупотреблений входит в корпус.

Следующей важной характеристикой корпуса является его репрезентативность. Корпус должен с максимальной объективностью представить разнообразие изучаемого явления и дать в то же время объективную картину бытования этого явления в речевой практике носителей данного языка.

Основным назначением ИКТ является использование его как информационной основы для получения количественных измерений (оценок) ЕЯ и для испытания лингвистических гипотез на различных структурных уровнях ЕЯ, начиная с алфавита и заканчивая текстом, и различных уровнях его глубины – от морфологии до семантики и прагматики.

Количественные измерения языка могут касаться:

а) комбинаций символов, морфем, канонических форм, словоупотреблений;

б) состава и комбинаций грамматических конструкций;

в) кодирования частей речи;

г) частотности лингвистических объектов разных структурных единиц языка как в тексте в целом, так и в отдельных его частях и т.д.

Лингвистические гипотезы могут, например, высказываться в отношении применимости формальных грамматических правил и ограничений на их использование; особенностей диалогового, учебного, научно-технического и других подъязыков; моделей выделения тех или иных единиц языка, классификации стилистических явлений, алгоритмов построения семантического пространства в языке и т.п.

Таким образом, ИКТ – это большая по объему автоматизированная макросистема, включающая ряд подсистем (корпусов) текстов как фондов, ориентированных на фиксацию фактов языка. Такие фонды относятся к подсистемам регистрирующего типа в отличие от систем, ориентированных уже на ту или иную интерпретацию языковых данных (фонды словарных статей, грамматик и т.п.).

Очевидно, что ИКТ сам по себе не может обеспечить решения всего аспекта задач, связанных с получением количественных оценок языка и испытанием лингвистических гипотез, особенно если речь заходит о более высоких, чем морфологический, уровнях его глубины. Он является лишь основой создания соответствующих средств в виде совокупности так называемого аннотированного корпуса текстов и инструментальных средств доступа, извлечения и анализа естественноязыковой информации.

Полезность корпуса возрастает, когда он аннотируется, т.е. каждое слово в нем снабжается лексико-грамматическим, синтаксическим или семантическим кодом в зависимости от уровня обработки текста. Аннотирование выполняется с учетом контекста. Аннотированный текст превращается в текст как хранилище лингвистической инфор-

мации. Для разных задач требуются различные уровни аннотации текста, на которых вырабатывается различный объем дополнительных сведений. Выделяются следующие уровни такого типа [8]:

а) лемматизированные тексты, в которых для каждого слова указывается его основная форма и часть речи;

б) тексты с морфологической информацией, в которых для каждого слова указываются его основная форма, часть речи и полный набор морфологических характеристик;

в) тексты с синтаксической информацией, в которых для каждого слова указываются его основная форма, часть речи и морфологические характеристики, а также для каждого предложения указывается его синтаксическая структура.

Аннотированный корпус применяется в максимально широком круге приложений. Для такого применения необходимо подобрать формат записи аннотационной информации, отвечающий следующим условиям:

а) наличие нескольких «слоев» информации, извлекаемых из разметки независимо друг от друга;

б) потенциальная расширяемость на типы информации, не охватываемые аннотацией на настоящем этапе.

Для аннотирования текстов на различных уровнях глубины языка используются определенные системы кодирования, преобразующие нечеткие лингвистические объекты в соответствии с некоторой единой процедурой в дискретные лингвистические единицы, что позволяет работать далее не с конкретными структурными единицами, а с их классами. Для этой цели служат специально разрабатываемые классификаторы, содержащие различные типы лингвистической информации [7]. На сегодняшний день главной в аннотированных корпусах текстов была и остается информация о частях речи, которая фиксируется в процессе лексико-грамматического кодирования, цель которого состоит в том, чтобы назначить каждой лексической единице код, указывающий на часть речи, или иначе лексико-грамматический код (ЛГК).

Таким образом, первым разрабатываемым классификатором (базовым) является лексико-грамматический классификатор. При его создании используется подразделение слов на лексико-грамматические классы, называемые традиционно частями речи. При этом учитывается, что если набор грамматических признаков, описывающих слово и составляющих его характеристику, представить в виде иерархической структуры, то высший ярус займет признак части речи, поскольку он покрывает практически всю лексику [9].

Синтаксический классификатор лексических единиц и отношений включает синтаксические классы (коды), которые используются для классификации структурных элементов синтаксически проанализированных предложений. Семантический классификатор лексических единиц и отношений содержит семантические классы (коды), которые используются для классификации структурных элементов дерева фразы на семантическом уровне.

Синтаксический и семантический классификаторы, во-первых, дополняют описанный выше лексико-грамматический классификатор и, во-вторых, классифицируют только те элементы и отношения, которые распознаются в текстовых документах на соответствующих уровнях глубины языка. Эти уровни определяют в конечном счете основные типы знаний, на извлечение которых ориентированы различные виды анализа текста.

Среди базовых лингвистических ресурсов особое место занимают машинные словари ЕЯ – упорядоченные конечные массивы лингвистической информации, представленные в виде списков или таблиц, удобных для размещения в памяти компью-

тера, и снабженные программами автоматического поиска и ведения. Машинный словарь всегда является одним из главных компонентов любой интеллектуальной информационной системы. Поэтому вопросы, связанные с организацией хранения словарей, поиска в них, корректировки и др., играют важнейшую роль при проектировании таких систем.

Особое место занимают аннотированные машинные словари, называемые базовыми или эталонными. Такой словарь включает максимально возможное количество слов ЕЯ, при этом каждому из них указано множество всех соответствующих ему вне контекста ЛГК.

Структурный анализ текстов позволяет выявить, что каждый ЕЯ функционирует в соответствии с единой и фиксированной системой правил. Такая система правил, иначе распознающих лингвистических моделей (РЛМ), является основой, необходимой для осуществления требуемого анализа текста. В частности, множество РЛМ для лексико-грамматического анализа, по сути, есть грамматика ЕЯ в ее классическом понимании.

РЛМ – это один из способов формализации языковой компетенции в целях автоматического анализа текста на всех уровнях его глубины. РЛМ применяются для эксплицитного описания конкретных языковых ситуаций и определенных действий над лексическими единицами, их свойствами, отношениями и т.п. в анализируемом тексте.

2. Технологическая схема создания базовых лингвистических ресурсов

Разработка промышленных информационных систем требует ориентации на огромные объемы реальных текстов, и используемые в этих системах лингвистические ресурсы должны соответствовать этим текстам, а значит, должны фактически извлекаться из этих текстов. Исходя из этого, можно определиться с принципиальной технологической схемой создания базовых лингвистических ресурсов.

Для получения такого рода ресурсов предлагается метод, основанный на иерархическом, в смысле уровня структурных единиц ЕЯ, представлении ИКТ так, что уровень r_0 (слов) определяет более высокий уровень r_0+1 (фраз), который в свою очередь определяет уровень еще более высокого порядка и т.д. вплоть до уровня r^*, соответствующего тексту в целом. Обозначим через $K^{(r_0)}$, $K^{(r_1)}$, …, $K^{(r^*)}$ классификаторы семантико-грамматических свойств структурных единиц языка соответствующего уровня, $A_L^{(r_0)}$, $A_L^{(r_1)}$, …, $A_L^{(r^*)}$ – процедуры семантико-грамматического анализа ИКТ, $T_L^{(r_0)}$, $T_L^{(r_1)}$, …, $T_L^{(r^*)}$ – корпусы аннотированных текстов для ИКТ. Применяя $A_L^{(r_0)}$ к ИКТ, получим $T_L^{(r_0)}$, затем применяем $A_L^{(r_1)}$ к $T_L^{(r_0)}$ и получаем $T_L^{(r_1)}$ и т.д., пока не получим $T_L^{(r^*)}$.

Каждый аннотированный корпус текстов $T_L^{(r\wedge)}$ представляет собой ИКТ с выделенными в нем для каждого текста структурными единицами ЕЯ L уровня $r\wedge$, для которых указаны их семантико-грамматические свойства в соответствии с классификатором $K^{(r\wedge)}$, сгенерированные процедурой $A_L^{(r\wedge)}$. В качестве структурных уровней языка рассматриваются уровни слова, фразы, предложения, текста. Процедура $A_L^{(r\wedge)}$ может быть ручной, автоматизированной или автоматической.

На начальном этапе вся работа сосредотачивается на уровне r_0, т.е. строится $T_L^{(r_0)}$ по следующей обобщенной схеме:

Этап 1. Разработать минимальный с точки зрения трудоемкости и точности решения планируемых задач ИКТ T_0.

Этап 2. Разработать классификатор $K^{(r_0)}$ семантико-грамматических свойств ЕЯ L на уровне слова.

Этап 3. Автоматически получить из T_0 исходный словарь словоформ ЕЯ L. Назовем его *эталонным словарем* и обозначим D.

Этап 4. В диалоговом режиме приписать каждому слову из D множество соответствующих ему вне контекста свойств (кодов) из $K^{(r_0)}$; полученный словарь обозначим $D^{(r_0)}$.

Этап 5. Используя словарь $D^{(r_0)}$, осуществить автоматическую идентификацию T_0: приписать каждому слову из T_0 соответствующее ему по словарю $D^{(r_0)}$ множество кодов.

Этап 6. В диалоговом режиме, используя контекст, снять, где необходимо, в «обогащенном» кодами на этапе 5 ИКТ T_0 многозначность.

В итоге формируется аннотированный корпус текстов ЕЯ L уровня r_0, в котором для каждого словоупотребления указан единственный код в соответствии с классификатором $K^{(r_0)}$. Отметим, что для последующих структурных уровней языка общая схема останется в целом такой же, только, например, на уровне фразы в качестве отдельной статьи словаря $D^{(r_1)}$ будет выступать цепочка кодов из $K^{(r_0)}$ и соответствующие ей коды из классификатора $K^{(r_1)}$ и т.д.

Ранее уже подчеркивалось, что при создании ИКТ первоочередной является задача определения его объема. И если первые корпусы текстов английского языка (The Lancaster/Oslo-Bergen Corpus (LOB), the Brown University Corpus), разработка которых велась уже в 60-е годы прошлого века, содержали 1 млн слов, то, к примеру, в корпусе современного английского языка The British National Corpus (BNC) 4124 текста с общим количеством слов более 100 млн [10]. При этом следует учитывать, что английский язык принадлежит к флективно бедным языкам, для которых различие между словом и словоформой (грамматической формой слова) практически отсутствует. Для флективно богатых языков, например, белорусского, русского, украинского, различие между словом и словоформой существенно. Слова в белорусском языке, в зависимости от части речи, могут иметь до 28 грамматических форм. Поэтому в идеале минимальный размер корпуса для таких языков, исходя из объема базового словаря в 100 тысяч слов, должен быть не менее 3 млн словоупотреблений.

Для определения репрезентативности ИКТ следует исходить из предположения, что ИКТ будет ориентирован на современные потоки научно-технической, деловой и общественно-политической информации, поэтому можно выделить следующие критерии подбора текстов:
– лексика должна отражать современное языковое употребление и представлять «основное лингвистическое поведение»;
– тексты должны быть широко читаемыми, поскольку именно такие тексты оказывают наибольшее влияние на развитие языка;
– могут не рассматриваться многие области технического языка за исключением тех, лексика которых «просачивается» в повседневное использование;
– исключаются маломасштабные области;
– вводятся ограничения на характеристики источников текстов и тематику.

При отборе текстов должны учитываться следующие основные положения:
– тексты должны быть жанрово однородны, т.е. принадлежать одному жанру, без цитирования произведений иных жанров;
– тексты должны быть авторски однородны (принадлежать одному автору), с минимизацией диалогов, цитирования иных авторов;
– тексты должны максимально отражать синтаксические и морфологические особенности жанра, т.е. при отборе текстов предпочтение должно отдаваться текстам с максимально выраженными различиями в пределах одного жанра (синтаксическими, стилистическими и т.д.);
– задачей ИКТ не является собственно анализ лексики, таким образом, тексты не должны быть перегружены малочастотной лексикой.

Аннотирование и разработка базового словаря требуют наличия классификаторов, позволяющих преобразовать все нечеткие лингвистические объекты в соответствии с некоторой единой процедурой в дискретные лингвистические единицы с использованием эффективной системы их кодирования [9].

Для ЕЯ различают такие структурные единицы, как морфема, словоформа, фраза, предложение, дискурс, текст. Каждая из этих структурных единиц образуется на основе определенных правил конкретного ЕЯ. В соответствии с этими правилами каждой j-й структурной единице i-го уровня, обозначим $t_j^{(i)}$, может быть поставлено в соответствие множество $s_j^{(i)}$ ее морфологических, синтаксических и семантических свойств, известное под названием *кода*.

В зависимости от поставленной задачи структурные единицы ЕЯ любого из указанных уровней с соответствующими множествами свойств могут быть взяты в качестве элементов словаря $(t_j^{(i)}, s_j^{(i)})$ (в самом общем смысле этого слова).

В настоящее время не существует стандартов на представление подобной информации, однако на основе проведенного анализа можно сделать следующие предположения относительно системы кодов:

а) код должен быть лаконичным и в то же время избыточным;

б) код должен снимать омонимию, т.е. быть однозначным;

в) код должен обеспечивать наличие нескольких «слоев» информации, извлекаемых из разметки независимо друг от друга, а также потенциальную расширяемость на типы информации, не охватываемые аннотированием на определенном этапе.

На уровне слова используется основной тип лингвистического кодирования – морфологический анализ, или аннотирование по частям речи. Он направлен на то, чтобы с минимальными потерями информации получить достоверное представление текстов различных предметных областей на уровне отдельных слов [11]. Подобный вид кодирования увеличивает определенность поиска данных в корпусах текстов и формирует основу для синтаксического и семантического анализа.

Перечисленные выше принципы описывают процесс формирования так называемых статических компонентов (собственно БД) базовых лингвистических ресурсов. Существует также и динамический компонент – БД распознающих лингвистических моделей.

В общем случае отдельную РЛМ формально можно представить в виде

$$\text{«условие} \rightarrow \text{операция»}. \qquad (1)$$

Условие представляет собой лингвистический паттерн (шаблон), задающий конкретную языковую закономерность. Паттерн – это формальная спецификация свойства набора примеров, определенная в терминах некоторого формального языка.

Приведем пример одного из возможных условий РЛМ:

«последовательность из двух слов, первое из которых является либо точкой, либо восклицательным, либо вопросительным знаком, а второе – словом с большой буквы».

Например, ... ***?*** *Ответом ...*

*... **.** Алгоритм ...*

*... **!** Это ...*

Следует заметить, что в данном конкретном примере языковая ситуация описывается с использованием только лексических единиц. Однако в общем случае, при ее описании могут использоваться и другие уровни языка: ЛГК, синтаксические и семантические классы и т.д.

Если некоторый фрагмент анализируемого текста удовлетворяет условию, то над ним производится соответствующая операция. Выделяется два основных типа операций в соответствии с тем, какие действия они проводят – оценочные операции (например, считать данный ЛГК у определенного слова истинным) и трансформационные операции (например, изменить ЛГК у конкретного слова из текста на другой ЛГК). Например, при выполнении приведенного выше условия операцией будет:

«считать первое слово, т.е. любой из трех указанных знаков, границей предложения»

и эта операция является операцией оценочного типа.

РЛМ указанного типа разрабатываются для задач форматирования текста, разбиения текста на слова и предложения, распознавания идиом, лексико-грамматического, основанного на правилах, анализа текста, синтаксического и семантического анализа текста. Распознающие лингвистические модели синтаксического анализа текста включают РЛМ распознавания именных и глагольных групп, а также глагольного управления; последние фактически реализуют начальный (базовый) этап уже семантического анализа текста.

3. Базовое лингвистическое обеспечение белорусского и русского языков

Описанные выше аспекты разработки базовых лингвистических ресурсов нашли свое отражение при построении Компьютерного фонда белорусского языка (КФБЯ) (работа осуществлялась научно-исследовательской лабораторией интеллектуальных информационных систем Белорусского государственного университета).

Был разработан классификатор лексико-грамматических свойств белорусского и русского языков, базовые словари и аннотированные корпусы текстов для указанных языков.

При разработке классификатора были учтены принципы кодирования применительно к языкам с разветвленной флективной системой. При создании классификатора использовалось подразделение слов на лексико-грамматические классы, называемые традиционно частями речи: имя существительное, имя прилагательное, глагол и др., исходя из того, что, если набор грамматических признаков, описывающих слово и составляющих его характеристику, представить в виде иерархической структуры, то высший ярус ее займет признак части речи, поскольку он покрывает практически всю лексику.

Далее учитывались не только классы слов, но и подклассы. Например, местоимения распадаются на ряд подклассов, различных по лексическим значениям, морфологическим формам и синтаксическим функциям, например, личные, возвратные и притяжательные местоимения и т.п.

Классификатор представляет собой систему грамматических свойств с элементами семантики белорусского языка и имеет два уровня. Первый уровень включает 96 кодов и характеризует словоизменительную парадигму в целом, т.е. это код, который одинаков у всех слов из парадигмы. Второй уровень содержит 63 кода и характеризует каждое слово в парадигме – словоформу, так как содержит грамматическую информацию. Совокупность кодов первого и второго уровней и образует уникальный код, который приписывается конкретной словоформе.

Были разработаны базовые словари для белорусского и русского языков. Общие размеры словарей составили: для белорусского языка – около 2,7 млн словоупотреблений, для русского – около 4 млн. Расчет производился для словоупотреблений, а не для слов, так как указанные языки принадлежат к флективным языкам и характеризуются богатой словоизменительной парадигмой.

Словарный состав всех указанных словарей постоянно обновляется.

Дадим характеристику некоторым из них.

Разработанный базовый словарь белорусского языка содержит слова, принадлежащие ко всем существующим в языке частям речи, а также вводные слова и причастия, и является словарем словоформ, сгруппированных в парадигмы – совокупности словоизменительных форм, представленных в памяти компьютера вместе с соответствующими им ЛГК.

В состав словаря имен собственных входят словники личных имен, фамилий и отчеств, наименований физико-географических объектов и территориальных единиц Беларуси, наименований мировых физико-географических объектов и территориальных единиц, других наименований (названия государственных и общественных организаций, религиозных праздников, литературных памятников, языков программирования и пр.). Эти слова повседневно употребительны, но, будучи именами собственными, традиционно в словари общей лексики не включаются. Данный словарь входит в многоярусную систему словарей языка, и без него описание сегодняшней белорусской лексики было бы неполным.

Словарь аббревиатур и сокращений содержит наиболее употребительные сокращения современного белорусского языка и призван показать систему сокращений как часть его лексического фонда.

Разработанный словарь синонимов белорусского языка содержит не только синонимы в их классическом понимании, но и варианты слова (например, *дзіця – дзіцё*), необходимые при информационном поиске и синтезе текста для их полного отождествления.

Словарь омонимов белорусского языка представляет собой словарь омоформ, поскольку лексические, или простые, омонимы содержатся в базовом словаре белорусского языка. В данном словаре не учитывается также внутрипарадигматическая омонимия, например, именительного и винительного падежей, если омонимичные формы содержатся в одной парадигме.

Разработанный словарь антонимов белорусского языка представляет собой не просто список противоположных по значению слов, а содержит синонимические ряды, которые между собой являются антонимичными (противоположными по значению).

Словарь ударений белорусского языка разрабатывался на основе базового словаря белорусского языка с указанием образования грамматических форм и особенностей расстановки ударения.

Важной составляющей лингвистического обеспечения КФБЯ являются терминологические словари – словари, содержащие терминологию одной или нескольких специальных областей знаний или деятельности. Их можно считать лингвистическими словарями подъязыков конкретных отраслей знания и/или видов профессиональной деятельности, тогда как с точки зрения общелитературного языка содержащаяся в них информация является скорее экстралингвистической. Число тематических словарей очень велико и постоянно увеличивается; многие терминологические словари к тому же являются двух- или многоязычными. Различные типы словарей предоставляют разные возможности для пользователя.

Особенностью разработанных терминологических словарей является их двуязычность: словари представлены на белорусском и русском языках. Данные словари могут служить основой для создания тезаурусов по предметным областям.

Далее, в соответствии с указанной принципиальной схемой, были разработаны:
– исходный корпус текстов для белорусского и русского языков (суммарный объем – 10 млн словоупотреблений);
– аннотированные корпусы текстов для белорусского и русского языков.

Общие размеры аннотированных корпусов составили: для белорусского языка – около 400 тыс. словоупотреблений, для русского – около 1 млн. Расчет производился также для словоупотреблений. Для расчета среднего количества кодов для каждого словоупотребления из рассмотрения были исключены знаки препинания, формулы и иностранные слова.

В качестве основного приложения КФБЯ была разработана информационная система, состоящая из словарей и аннотированного корпуса текстов, которая служит для информационно-справочного обслуживания пользователей относительно современного белорусского языка в его письменной форме и обеспечения доступа к лингвистическим компонентам фонда. Словари доступны пользователю как справочное средство (поиск слов, предоставление информации относительно словоизменения конкретных единиц словаря).

Заключение

Разработка лингвистических ресурсов – достаточно долгий и трудоемкий процесс, требующий привлечения высокопрофессиональных экспертов в области языка и соответствующих приложений одновременно.

Предложенная технология разработки базовых лингвистических ресурсов ЕЯ является универсальной, что и было продемонстрировано при разработке лингвистических ресурсов белорусского языка.

Эти ресурсы используются в различных системах, выполняющих обработку ЕЯ: корректировки орфографии, машинного перевода, автоматического реферирования, информационного поиска и других.

Предлагаемый подход может быть использован в аналогичных разработках для других естественных языков.

Литература

1. Карпов В.А. Язык как система. – Минск: Вышэйшая школа, 1992. – 302 с.
2. Логический подход к искусственному интеллекту: от модальной логики к логике баз данных: Пер. с франц. / Тейз А., Грибомон П., Юлен Г. и др. – М.: Мир, 1998. – 494 с.
3. Пиотровский Р.Г. Автоматическая переработка текста: теория и практика к концу XX в. // Научно-техническая информация. – Сер. 2. – 1998. – № 5. – С. 26-36.
4. Roubashko N.K. Development and Representation of Linguistic Knowledge for Natural Language Processing // Systems and Signals in Intelligent Technologies (SSIT'98): Proc. of the International Conference. – Minsk. – 1998. – P. 383-389.

5. Лейкинд Ю.Е. Архитектура современных лингвистических процессоров и их информационные ресурсы // Информационные системы и технологии (IST’2002): Материалы I Междунар. конф.: В 2 ч. – Минск. – 2002. – Ч. 1. – С. 153–158.
6. Linguistic Data Consortium – Mode of access: http://www.ldc.upenn.edu/
7. Mcenery T., Wilson A. Corpus Linguistics. – Edinburgh: Edinburgh University Press, 1996. – 132 p.
8. Богуславский И.М., Григорьев Н.В., Григорьева С.А. и др. Аннотированный корпус русских текстов: концепция, инструменты разметки, типы информации // Труды Международного семинара Диалог’2000 по компьютерной лингвистике и ее приложениям. – Протвино. – 2000. – Том 2. – С. 41-47.
9. Рубашко Н.К. Разработка лексико-грамматического классификатора флективных языков // Вестник МГЛУ. Сер. 1: Филология. – 2003. – № 11. – С. 84-89.
10. The British National Corpus. – Mode of access: http://www.heu.ox.ac.uk/BNC/
11. Совпель И.В. Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста. – Минск: Вышэйшая школа, 1991. – 118 с.
12. Рубашко Н.К, Невмержицкая Г.П., Совпель И.В. Компьютерный фонд белорусского языка и его приложения // Информационные системы и технологии (IST’2006): Материалы III Междунар. конф.: В 2 ч. – Белорус. гос. ун-т. НАН Беларуси. Науч.-техн. ассоциация «Инфопарк», Акад. упр. при Президенте Респ. Беларусь. – Минск. – 2006. – Ч. 2. – С. 71-76.

Н.К Рубашко

Розробка базових лінгвістичних ресурсів природної мови для інформаційних систем

У статті викладаються основні принципи розробки базових лінгвістичних ресурсів природної мови. Наводиться перелік основних складників, надається опис технології створення цих ресурсів. У якості прикладу використовування наданої технології розглядається розробка базових лінгвістичних ресурсів білоруської та російської мови.

N.K. Roubashko

Development of Natural Language Basic Linguistic Resources for Information Systems

The article deals with the main principles of natural language basic linguistic resources development. The basic components and resources creation technology are described. The development of basic linguistic resources of the Belarusian and Russian languages is given as a usage example of suggested technology.

Статья поступила в редакцию 17.07.2008.